Technical Realization

We have implemented the different ML models for each scenario described in Herm et al. (2021). For the implementation, we use the python package scikit-learn as well as keras. Further, each result presentation is set in the same color scheme. We selected the parameters through hyperparameter tuning using scikit-learn's GridSearch. All corresponding technical implementations are described below.

Dataset Preprocessing. For the data preprocessing, we use the recommended steps according to García et al. (2016). Therefore, we first deleted any incomplete observations as well as outliers. Afterward, we performed a feature selection. Additionally, for the Bakery dataset, features such as special promotions were combined. Subsequently, we applied a min-max normalization to minimize variations.

Linear Regression. We applied a default configuration of this model given by the scikit-learn package.

Decision Tree. We did not restrict the model by any parameters such as maximum depth or features to maintain the highest possible generalizability for the implementation of the decision tree.

SVM. We implemented an SVM with a radial basis function (RBF) as kernel function. We have not implemented feature importance, since the rbf-kernel function does not allow for the calculation.

Ensemble Learning. We used the bagging algorithm random forest, due to its frequent usage in practice (Lundberg et al., 2020). We did not restrict the random forest by any parameters. Since an ensemble algorithm consists of multiple weak estimators, we used 100 estimators.

ANN. For Bakery and WINE, we apply a multi-layer perceptron. Due to the time series data in the C-MAPSS dataset, we use an LSTM in this case. All models use the ReLU activation function, dropout layers as well as early stopping.

XANN. Since our literature review revealed an extensive application of the XAI augmentation SHapley Additive exPlanations (SHAP), we used this python package to transfer our ANNs into a grey-box model. We trained the explainer kernel using the full training set from the datasets. Global feature impact as well as a local explanation were applied.

Performance of Models

Table 1 shows the performance of the AI models for the three scenarios. Overall, the results fit to the theoretical ranking presented in Herm et al. (2021). Furthermore, we show the mean of the inverted normalized RMSE to enable a ranking as well as a comparison within the different models' overall scenarios. Please note that ANN and XANN use the same AI model.

Table 1. Result of Performance Measurement

	Performance in RMSE			
Models	**WINE***	**Bakery Chain***	**C-MAPSS***	**Normalized Mean****
XANN	0.68	141.41	36.13	**1.00**
ANN	0.68	141.41	36.13	**1.00**
Ensemble Learning	0.72	144.21	38.27	0.91
Decision Tree	0.79	156.50	58.74	0.62
SVM	0.74	227.65	53.03	0.50
Linear Regression	0.89	245.06	83.18	0.00

* lower is better, ** mean of inverted normalized RMSE over all scenarios

As assumed in the theoretical classifications, there is no clear gap between the performance of decision trees and SVMs. For the WINE and C-MAPSS scenario, SVM performs better while in the Bakery Chain scenario, SVM's results are by far the worst. Furthermore, according to the theoretical assumption of Figure 1, ANN and XANN archive the lowest RMSE in every scenario. Likewise, the ensemble learning model random forest archive the second lowest RMSE results.

References

Herm L.-V., Wanner J., Seubert F., Janiesch C. (2021). I don't get it, but it seems valid! The Connection Between Explainability and Comprehensibility in (X)AI Research. *In: 29th European Conference on Information Systems (ECIS). AIS Virtual Conference Series* : AIS, 2021, bl conditionally accepted.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. and Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, 1 (9), 1-22.